Guidance for Handling Different Types of Data

March 2018

Introduction

This guide introduces the different types of data that may be collected as part of an activity and includes suggested formats for numeric data, coding for text data and suggestions on storing images, video clips, and audio files.

Common Data Types

This section of the guide details the most common types of data that are usually stored as variables, that is: identifier, numeric, categorical, date, and some free text data. It also contains advice about things to consider for each type of data, and what additional information needs to be collated and archived alongside the main datasets.

Figure 1 shows some data that have been compiled poorly and contains several common data mistakes most of which originated at the stage of designing the data collection tool and lack of guidance to the enumerators and data entry staff. The contents can mostly be understood by examining the data by eye (though this would not be the case for a large dataset); but a large amount of data manipulation would have to be conducted for the data to be used in any statistical software package. While you read through this section, consider how this information should have been recorded; Figure 5 at the end of this section, presents one possible "good" option for storing the same data.

Figure 1 - Example of poor data storage/collection

	A	B	C	D	E	F	G	H	I	J	K
1											
2		Identifier	Sex	Date of Birth	Size of Household	Crops	Farm Area	2008 Yield (t/ha)	Yield 2009 (t/ha)	% Change	
3						Grown				*2008-09*	
4		1	Male	28-Feb-1965	6	Maize; Sorghum; Cassava	1 h	2	3.452	75%	
5		D002	M	3/3/54	5	Maize	1.5 acre	2.1	2	5%	
6		D003	f	1970-01-03	4.5	Sorghum; Cassava	0.6666666667 hect	6	6.45	8%	
7		4	F	March 12th 1969	Unknown	maize	2	1	1.658	65.80%	
8		5	male	5-Oct-1945	2	Maize	1 acre	2	2.463	23.2%	
9		D006	M	12/12/81	1	Only sorghum	2 h	3.67	4	10%	
10											
11							Means of yields:	2.795	3.337166667	0.311667	
12											

Identifier Data

The identifier data for an activity are often referred to as the unique identifier or primary key. All datasets must have a variable or set of variables that uniquely identifies each record. For example, in a household survey, the region, village and household number are all identifier data and often these variables would be combined to create a primary key that uniquely identifies each household in the survey.

Identifier variables can be numeric, text, or alphanumeric depending on the context. Generally, codes are used so that the data can easily be anonymised; using an individual's name as the primary key is not recommended as more than one individual might have the same name. Datasets should contain as many levels of identifier variables in separate columns as appropriate, for example, in a household survey, the household level data should contain the household ID and any data at the level of individuals within the households should contain both the household ID and the individual ID; including the household ID at the individual level is essential as without this we would be unable to link the individual level data with the household level data.

Numeric/Continuous Data

Numeric or continuous data are the result of questions or measurements that naturally give a numeric response. This could be the number of household members, yield measurements, plot size, age, amount of fertiliser applied, etc.

Different types of numeric data will require slightly different formats to account for different levels of precision. For example, age is generally recorded as the number of completed years with no decimal places or fractions; on the other hand, when measuring a response such as yield, is it usually appropriate to include a few decimal places – whether 1, 2 or 3 decimal places are required is dependent on the measurement device and the intended use of the data. When collecting numeric data, it is important to choose the format in advance and make all enumerators aware of this format to avoid rounding errors. For example, in datasets with height measurements, where the height of the individual was to be recorded to the nearest millimetre, you will often find peaks at half and whole centimetres. When specifying formats for numeric data it is important to be consistent with the precision within each variable.

When numeric data are the result of measurements, the units should be specified in advance or be recorded in a separate text column if they are not consistent across all observations (e.g. plot sizes in hectares, acres and local units).

Thought needs to be given about how missing values should be specified; for categorical data it is common to use '99' to indicate a missing response, however, when the data are numeric, this could be a plausible value. Similarly avoid using 0 to represent missing values. Alternatives are to leave the response as a blank (although with blanks it is difficult to distinguish between missing values and data that hasn't yet been entered), or use an impossible value, for example missing values for age and yield could be recorded as -99. Where numeric variables could reasonably take both positive and negative values, you would need to ensure that the missing value codes are a factor of 10 larger than the highest expected value.

Additional numeric variables can be derived as part of the transition from raw to primary data, for example converting areas recorded in different units into one common unit or calculating percentage change over time. Derived variables should not be calculated at the time of data collection; this is to avoid errors in the data and to ensure quality. The syntax used for deriving these variables should be documented and archived with the data.

Categorical Data

Categorical data, or factors, are questionnaire or measurement responses that fall into categories. For example, respondent gender: male/female, plot attacked by pests: yes/no, highest level of education, etc.

Categorical variables can be treated as either restricted text variables or restricted numeric variables. If they are being stored as text, the levels should be pre-specified and entered identically across all instances; for example, gender could be recorded as 'Male' or 'Female'. Remember, though that most analytical software is case sensitive and will therefore treat 'Male' and 'male' as two separate levels so it is important that the text entered at the data entry or collection stage is identical.

Data Dictionary / Code Lists

To avoid problems with different spellings or cases in the text, numeric codes are generally used. Each category should be assigned a code and an accompanying data dictionary must always be provided to decode the information. For example, 1=Male, 2=Female. Code lists (from within a data dictionary) can be recycled across questions if the same categories are used in more than one question, for example, with a single list of occupations, both the primary and secondary occupations of an individual can be recorded.

When producing a data dictionary and the code lists, spend time ensuring that all the lists contain all relevant options to catch potential missing values; for example, when asking about the highest level of education, include a code for ‘no formal education’ so that these responses are not classed as missing along with the true missing values.

Using numeric codes for categorical data is generally more efficient and less prone to errors. Value labels should be included in the dataset when it is analysed so, although the response was entered as 1 and 2, the software will use the value labels and show ‘Male’ and ‘Female’ in the tabulations and graphical output.

When storing categorical data as numeric codes, it is vital to make clear what each code represents. When producing the data collection tool, the codes for the categorical responses should be clear for the enumerators; see Figure 2 for an example.

Figure 2 - Illustration of coding categorical responses in the data collection tool

B1: Household composition and basic data. Include all members of the household who live in the dwelling and usually eat meals together. Include those who are temporarily absent (less than 6 months in the 12 months prior to the interview date). Do not include guests or paid workers. Use another separate sheet, if required.

A	B	C	D	E	F	G	H	I	J	K
S#	First Name	Sex 1. **Male** **2.Female**	Age (completed years) **999. Do not know**	Marital status **(see code)**	Relation to head of household **(see code)**	Currently in an educational institution **(see code)**	Adult literacy level *Only ask those 15 yrs or older* **(see code)**	Highest level of education completed **(see code)**	Main Occupation **(see code)**	Second Occupation **(see code)**
		|___|	|___|___|___|	|___|	|___|___|	|___|___|	|___|	|___|___|	|___|___|	|___|___|
		|___|	|___|___|___|	|___|	|___|___|	|___|___|	|___|	|___|___|	|___|___|	|___|___|
		|___|	|___|___|___|	|___|	|___|___|	|___|___|	|___|	|___|___|	|___|___|	|___|___|
		|___|	|___|___|___|	|___|	|___|___|	|___|___|	|___|	|___|___|	|___|___|	|___|___|
		|___|	|___|___|___|	|___|	|___|___|	|___|___|	|___|	|___|___|	|___|___|	|___|___|
		|___|	|___|___|___|	|___|	|___|___|	|___|___|	|___|	|___|___|	|___|___|	|___|___|
		|___|	|___|___|___|	|___|	|___|___|	|___|___|	|___|	|___|___|	|___|___|	|___|___|
		|___|	|___|___|___|	|___|	|___|___|	|___|___|	|___|	|___|___|	|___|___|	|___|___|
		|___|	|___|___|___|	|___|	|___|___|	|___|___|	|___|	|___|___|	|___|___|	|___|___|

Codes for Column E	Codes for Column F	Codes for Column G	Codes for Column H	Codes for Column I	Codes for Columns J and K
1. Married 2. Divorced 3. Separated 4. Widowed 5. Single	1. Head 2. Spouse 3. Mother/father 4. Sister/brother 5. Son/daughter 6. Grandson/granddaughter 7. Daughter-in-law / son-in-law 8. Nephew/niece 9. Aunts/uncles 10. Cousins 11. Brother-in-law/Sister-in-law 12. Mother-in law / Father-in-law 98. Other _________	1. Govt School 2. Private school 3. Madrassa 4. Govt school and Madrassa 5. Private school and Madrassa 6 Adult Literacy 7. Kindergarten 8. Early Child Development (ECD) Centre 9. At University 10. Not in school	1. Cannot read and write 2. Can read only 3. Can read and write 9. Not applicable (for those aged <15 yrs)	0. No education For grade 1-14 record the actual grade completed 15. Technical/ Vocational Graduate 16. University Graduate 17. Madrassa only 98. Other ___________ 99. Do not know	0. No occupation 1. Private sector employee (including NGOs) 2. Government employee 3. Daily wage earner (casual worker) 4. Self employed (Trade / Business) 5. Livestock rearing 6. Crop Farming 7. Fishing 8. Unemployed- job seeking 9. Housework 10. Student 11. Retired/Pensioner 12. Ill/disabled 98. Other _______________

The information in the code lists should be collated across all sections or questions in the tool to form the data dictionary which should be stored as part of the metadata. The data dictionary can provide useful information concerning the exact contents of each variable; this includes numeric variables for which the measurement units and codes for N/A (not applicable) or missing values can be specified. SPSS, CSPro, Stata and other packages support the creation of data dictionaries

particularly well, automatically producing the data dictionary alongside the main dataset, outlining column information and the values of all the codes.

Dates

Dates should be unambiguously recorded so that there is no confusion over the ordering of the day, month and year at any stage of the process. We suggest the use of separate columns for day, month and year, then converting these into a standard date format during the transition from raw to primary data. This consistent approach is particularly important in situations where work in conducted across countries with different calendars or different notations for recording dates.

When using dates, it is important to consider the software packages being used to enter and analyse the data. Some packages, such as SPSS, have specific date formats, while others, such as CSPro store dates as strings of digits. Transferring dates from one package to another is often not as straightforward as it should be, and dates are often corrupted during transfer; care needs to be taken to ensure quality.

Free Text Data

Free text data is text data that cannot be coded into categories. This includes items such as name of respondent or data resulting from open ended questions such as 'give your opinion on …', 'do you have any further comments …', etc.

Free text should be entered exactly as seen in the original source; no text should be paraphrased. Take care when entering non-ASCII text characters and ensure your software can cope with these characters. Where possible free text should avoid punctuation which may corrupt the data when input into analytical software (e.g. ; , %) due to conflicts with the definitions of the characters in such software. For example, using commas in free text may cause problems if you are using csv (comma separated variables) format as data may end up in the wrong column.

If free text is needed for any analysis, for example if the frequency of respondents mentioning a key phrase or word is of interest, it may be useful to enter the data either all in upper case or all in lower case.

Missing Data

> *"There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know. And if one looks throughout history … it is the latter category that tends to be the difficult one."*
>
> *Donald Rumsfeld, US Secretary of Defence, 2002*

All projects should strive for 100% of data to be of the 'known known' variety, but practically this is rarely possible. However, it is always possible to make sure that your activity avoids the problem of 'unknown unknowns' by recording and retaining information about why data are missing. Where possible the study should be designed to allow for as much information to be collected as possible. Allow for options such as 'other, please specify', and 'N/A' responses to be specified as well as

comment variables so that enumerators can record additional information where necessary about why certain data are missing.

Except for data collected from clinical trials, there are no fixed guidelines for the storage of missing data and almost every software package recognises and treats missing values in a slightly different way. Therefore, the most important aspect of missing data is within study consistency and ensuring that the processes are well documented.

When designing data collection tools, it is important to enable the enumerator to distinguish between informative non-responses and missing data; you will have to train the enumerators to understand the distinction with respect to your specific activity. For example, when measuring yield, if the measurement could not physically be taken as weather conditions restricted enumerators from visiting the plot, then a missing value should be recorded. However, if the entire plot has been destroyed by pests, even though no measurement was taken, the yield should be recorded as zero rather than a missing value. Informative non-responses may also occur in questions listing a series of conditions and requiring the responder to 'tick all that apply'. By implication any boxes left un-ticked indicate a response of "no" and should be treated as such. However, with this style of question, there is no distinction between 'no', 'N/A' and missing. Adding an option for "None of the above" will help with this problem.

When assigning missing value codes to an activity tool, such as a questionnaire or field report, it should be possible to consistently apply the same code to indicate missing data across all variables regardless of their data type (apart from date variables). If the enumerator is consistently using the same code, whether this is 99, -99, 999 or -999, to indicate a missing response for the whole data collection activity, they are far less likely to make mistakes when recording missing values.

There are many different scenarios where missing data could occur and to capture these reasons it would be necessary to provide different codes to explain why data are missing. Suggestions for codes and examples of missing data types are given in Table 1. This is presented as a guideline only, if certain missing data scenarios are not of interest then the number of codes can be reduced; if there are other activity specific missing data scenarios, then the number of codes can be increased.

Table 1 - Missing value codes

Missing Data Type	Example	Suggested Code
Unknown	Responder does not know the answer	-99
Not applicable	Question is not relevant for the circumstances given previous responses	-98
Non response	Responder refused to answer the question or allow measurement to be taken	-97
Non agreement	If more than one responder was present and they could not reach a consensus on the correct response	-96
Not recorded / Human error	Variable was left blank mistakenly or measurement was not taken	-95
Technical error	Measurement equipment failed to provide valid output and could not be repeated	-94

Multiple Response Data

Multiple response data are usually the result of responses being a list, for example:

- Name the main crops grown on your land;
- What livestock do you keep?
- What fertiliser(s) have you applied to this plot?
- Etc.

Multiple response data can be managed in two possible ways and thought needs to be given to this when designing the data entry software (or setting up the worksheets in software such as Microsoft Excel in which the data are to be entered):

- The number of variables (columns) that are dedicated to a multiple response question should be the maximum number of responses given by any individual. The first response given by a respondent is entered into the first variable, the second response into the second variable, etc. If a respondent only gives two responses then only the first two variables will contain data, the other variables should be left blank. Figure 3 is based on the multiple response question of: "Name the main crops grown on your land". The third individual has listed 6 crops; this is the maximum number of crops listed by any individual therefore six crop variables have been created to store the data relating to this question. The first individual has only specified Maize and Sorghum, so the remaining crop variables have been left blank.

Figure 3 - Option 1 for storing multiple response data

	A	B	C	D	E	F	G
1	**IndID**	**Crop1**	**Crop2**	**Crop3**	**Crop4**	**Crop5**	**Crop6**
2	1	Maize	Sorghum				
3	2	Sorghum	Maize	Cassava			
4	3	Rice	Sesame	Cassava	Maize	Sorghum	Groundnut
5	4	Maize					
6	5	Sorghum	Rice	Cassava			
7	6	Maize	Sorghum	Sesame			

With this method of entering multiple response data, you cannot determine how many variables you will need until all the data have been collected. This would cause problems if you are using digital data collection where you would need to create your data entry system in advance. One solution would be to ask for up to a specific number of crops; for example: "what are your three most important crops?"; you would then create the data entry system with three variables for crops.

- The second option for storing responses to a multiple response question is to create a variable for each unique response given and indicate whether each crop was specified by the respondent using Yes/No or code 1/0 responses. Figure 4 shows the same data as Figure 3 but laid out using this second option.

Figure 4 - Option 2 for storing multiple response data

	A	B	C	D	E	F	G
1	IndID	Maize	Sorghum	Cassava	Rice	Sesame	Groundnut
2	1	Y	Y	N	N	N	N
3	2	Y	Y	Y	N	N	N
4	3	Y	Y	Y	Y	Y	Y
5	4	Y	N	N	N	N	N
6	5	N	Y	Y	Y	N	N
7	6	Y	Y	N	N	Y	N

The second option is preferable from an analysis point of view, due to the analysis of such variables being more straightforward; for example, to see how many respondents group maize there is only one variable to consider. However, in situations where respondents could specify a large range of different options, this second option may not be the most sensible as it could result in a large number of variables and only one or two respondents being flagged as 'Yes' for each. Some of the less popular crops may only have one or two respondents with a 'Yes' response.

Option 2 requires us to know which crops have been mentioned by any respondent before we can create the data entry system. Again, this causes problems with digital data collection where you need to have the data entry system designed before data collection. The solution is to determine in advance which crops you are interested in and phrase multiple questions as "Do you grow maize?", "Do you grow sorghum?", etc.

Better data storage

We return now to the data we saw at the start of this guide in Figure 1. Figure 5, shown below, shows the same data as Figure 1, but it has been stored more sensibly using the general advice given in this section. The ID variable is consistent as is the date of birth and the level of precision in the area variable; the area units are stored in a separate text variable and are all in upper case and spelt correctly. The crop variables are stored as multiple responses using option 2 above (with 1=Yes, 0=No). The yield data for the different years is laid out in a vertical format (all the yields in one variable, and another specifying the years) and stored on a separate sheet; this format enables analysis to be conducted more easily.

Figure 5 - Example of a better alternative data storage/collection (updated from example in Figure 1)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	SUBJID	GROUPID	SEX	DOB	HHSIZE	MAIZE	SORGHUM	CASSAVA	RAWAREA	RAWAREAUNIT	AREACONVERT	AREA	AREAUNIT
2	D001	D	M	1965-02-03	6	1	1	1	1.0	HECTARE	1.000	1.0	HECTARE
3	D002	D	M	1954-03-03	5	1	0	0	1.5	ACRE	0.404	0.6	HECTARE
4	D003	D	F	1970-01-03	5	0	1	1	0.7	HECTARE	1.000	0.7	HECTARE
5	D004	D	F	1969-03-12	-98	1	0	0	2.0	HECTARE	1.000	2.0	HECTARE
6	D005	D	M	1945-10-05	2	1	0	0	1.0	ACRE	0.404	0.4	HECTARE
7	D006	D	M	1981-12-12	1	0	1	0	2.0	HECTARE	1.000	2.0	HECTARE
8													
9													
10													
11													
12													

Demography | Yield

	A	B	C	D	E	F
1	SUBJID	GROUPID	YLDYEAR	YLD	YLDUNIT	YLDPERCHANGE
2	D001	D	2008	2.00	t/ha	
3	D002	D	2008	2.10	t/ha	
4	D003	D	2008	6.00	t/ha	
5	D004	D	2008	1.00	t/ha	
6	D005	D	2008	2.00	t/ha	
7	D006	D	2008	3.67	t/ha	
8	D001	D	2009	3.45	t/ha	72.6
9	D002	D	2009	2.00	t/ha	-4.8
10	D003	D	2009	6.45	t/ha	7.5
11	D004	D	2009	1.66	t/ha	65.8
12	D005	D	2009	2.46	t/ha	23.2
13	D006	D	2009	4.00	t/ha	9
14						

Demography | Yield | Sheet3

The types of data described in this section cover most of the data that are collected as part of a research activity. The remainder of this document goes onto give general advice about how to manage data and layout datasets/worksheets, as well as how to store other types of data such as images and videos.

General Variable / Dataset Specifications

- Each cell in a worksheet/database should only contain a single piece of information. For example, see the 'Crops grown' variable in Figure 1, data such as this should be stored as multiple response data. Similarly, a measurement such as yield should be stored in one variable and the measurement unit (if not pre-specified) should be stored in another variable.
- Variable names should always be included in the first row of any worksheet. These names should be descriptive and concise, but full details of the column contents should be included in the metadata rather than as part of the dataset. To prevent problems when using the data with analytical software, these names should contain no spaces or punctuation and should not begin with a number. In the past it was necessary to restrict variable names to 8 characters, while this is no longer a restriction for most software, it is still sensible not to make variable names too long. Consistent usage of capitalisation, always upper or lower case, is not a requirement but often makes analysis easier in software which is case sensitive. Variables which appear in more than one dataset/worksheet should always be named consistently. No two variables providing different information should be given the same name.
- When the same data are repeatedly collected over time, or as replicates, it can be easier at the analysis stage if these data are entered in a vertical format. Each additional time point or replicate is added to the dataset as an extra row with a variable indicating the visit or replicate number; see the second image in Figure 5 for an example of data in a vertical format.

- In many activities some information is collected only once but other information is collected on multiple occasions. For example, demographic variables are generally collected at the start of an activity, whereas yield information is collected each season. Here it becomes more efficient to store the demographic variables in a separate dataset/worksheet from the yield variables. This is an example of data being collected at multiple levels. Another example of data at multiple levels is collecting household level data such as household asset ownership, house building material, household size, etc., then collecting information on individuals within the household such as savings, group membership, etc.
- All information in the datasets/worksheets should be conveyed using numbers or text and not through formatting such as modifying text style or shading cells.
- Finally, datasets/worksheets should not contain any extraneous information, for example summary statistics or general data comments. Summary statistics should be included in the analysis and should not be stored as part of the data. Any general comments can be added to the metadata.

Other Data Formats

Data are not just items which can easily be input into spreadsheets. Data encompasses study document, reports, protocols, photographs, automated computer output, videos, maps, technical drawings, presentations, etc.

Documents

Please refer to the guides on Data and Document Storage and on Archiving for details of how to store documents such as protocols, data management plans and analysis reports.

Images

Images, including photographs, maps, graphs and technical drawings, should be retained in their original raw format on file. When placed into the activity archive they should be saved in an uncompressed, non-proprietary format. Depending on the size of the image, the best options currently are either .png (preferred) or .jpeg.

Audio / Video

All audio and video files should be retained in their original format on file. When placed into the archive they should be saved in an uncompressed, non-proprietary format. However, uncompressed video files are often extremely large, so for instances where many large files need to be archived, compression may be necessary.

Due to the continually evolving formats for digital audio and video, many storage formats have short life-spans as better compression algorithms are developed. The following guidelines for audio and video storage formats are taken from 2011 recommendations made by the California Digital Library - https://www.cdlib.org/gateways/docs/cdl_dffr.pdf. These are reviewed and updated from time to time.

Audio

We recommend using one of the following formats:

- Broadcast WAVE Audio File Format (BWF or BWAV)
  - Uncompressed linear pulse-code modulation (PCM)
  - 96 or 48 kHz/24 bits
- Audio Interchange File Format (AIFF)
  - Uncompressed linear pulse-code modulation (PCM)
  - 96 or 48 kHz/24 bits

Video

Optimal (uncompressed)

- Material eXchange Format (MXF) container format
  - Uncompressed YCbCr or JPEG2000 lossloss encoding (codecs). (In general, we recommend selecting a codec that is broadly adopted and well-documented, supported by multiple systems and vendors, and when possible, open-source)
  - 640 x 480 resolution (assuming 4:3 original aspect ratio)
  - 4:4:4 sampling scheme
  - 30-bit sample size
  - Progressive scanning
  - 30 MBps data rate

Alternative minimum (compressed)

- Audio Video Interleave (AVI) or QuickTime (MOV) container format
  - H.264/MPEG-4 AVC or DV encoding (codecs). (In general, we recommend selecting a codec that is broadly adopted and well-documented, supported by multiple systems and vendors, and when possible, open-source)
  - 4:2:2 sampling scheme
  - 30-bit sample size
  - Progressive scanning
  - 30 MBps data rate

Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at https://www.youtube.com/channel/UCs7EU95YMjhvNozJKCD92xQ/playlists. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes a video on “Guidance for handing different types of data” which is available at:
https://www.youtube.com/watch?v=SrRN2eHOVxk&index=16&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj